A SUPPORTING DISCUSSIONS

A.1 CONCRETE EXAMPLES OF GENERIC GENERALIZATION BOUND

- when **A1** is "Θ is finite, $l(\cdot,\cdot)$ is a zero-one loss, samples are *i.i.d*", $\phi(|\Theta|, n, \delta) = \sqrt{(\log(|\Theta|) + \log(1/\delta))/2n}$
- when **A1** is "samples are *i.i.d*", $\phi(|\Theta|, n, \delta) = 2\mathcal{R}(\mathcal{L}) + \sqrt{(\log 1/\delta)/2n}$, where $\mathcal{R}(\mathcal{L})$ stands for Rademacher complexity and $\mathcal{L} = \{l_\theta \,|\, \theta \in \Theta\}$, where l_θ is the loss function corresponding to θ.

For more information or more concrete examples of the generic term, one can refer to relevant textbooks such as (Bousquet et al., 2003).

A.2 ESTIMATION OF $c(\theta)$

The estimation of $c(\theta)$ mainly involves two difficulties: the knowledge of f_p and the computational cost of the search over the entire space $\mathcal{X}$. The first difficulty is usually resolved with intuition or common sense of the data or the task: in practice, we usually directly have the knowledge of $\mathcal{A}(f_p, \mathbf{x})$, *i.e.*, the spuriously correlated features that f_p relies on, such as texture of images. Therefore, the estimation becomes a process to test the whether the model will switch its correct prediction when these features are perturbed over the possible space. The second difficulty can be alleviated due to the fact that the search can be terminated once $r(\theta, \mathcal{A}(f, \mathbf{x}))$ is evaluated as 1. As one may be aware of, this process of searching the entire space with perturbations allowed in a predefined scope to test the model's worst possible prediction for a sample $\mathbf{x}$ is widely known as adversarial attack (Goodfellow et al., 2015). These techniques also usually leverage the knowledge of the model's gradient to accelerate the searching process.

While adversarial attack can offer a fairly accurate estimation of $c(\theta)$, it usually requires heavy computational efforts. As an alternative strategy, many other literature have tested the models with some fixed perturbations of the $\mathbf{x}$, or in other words, taking advantage of the fact that

$$|\theta(\mathbf{x}') - \mathbf{y}| \leq \max_{\mathbf{x}_{\mathcal{A}(f,\mathbf{x})} in \mathcal{X}_{\mathcal{A}(f,\mathbf{x})}} |\theta(\mathbf{x}) - \mathbf{y}| = r(\theta, \mathcal{A}(f, \mathbf{x})), \quad \text{where} \quad \mathbf{x}'_{\mathcal{A}(f,\mathbf{x})} \in \mathcal{X}_{\mathcal{A}(f,\mathbf{x})}. \tag{17}$$

to test a lower bound of $c(\theta)$. There are many works in this thread, and we only list a handful of examples: Jo & Bengio (2017) leveraged Fourier transform to show that models can capture a significant amount of texture information, later Geirhos et al. (2019) showed that CNNs trained with ImageNet are also biased towards texture. With a more concrete definition of the texture, Wang et al. (2020) demonstrated the models can capture high-frequency signals from images, which also links the discussion of learning through bias signals to the adversarial vulnerability issue of models (Ilyas et al., 2019). Similarly, these works mostly depend on a subjective choice of $\mathcal{A}(f_p, \mathbf{x})$, usually given by the knowledge of the data or the task. Although these works did not directly assess $c(\theta)$, θ usually switched the prediction for sufficient samples to raise an alarm.

A.3 LEARNING ROBUST MODELS WITH MINIMUM SUPERVISION IN PRACTICE

In practice, as we do not have the knowledge of either f_d or f_p ($\mathcal{F}$), the strategy we use is to estimate the model first and consider our estimated model $\widehat{\theta}$ as a substitute of the labeling function (either f_d or f_p). Therefore, at each iteration t, we will use the $\widehat{\theta}$ at the previous iteration to identify the active set for the optimization of (16) (in main manuscript).

Further, another question is that when we have $\widehat{\theta}^{t-1}$, how to identify $\mathcal{A}(\widehat{\theta}^{t-1}, \mathbf{x})$, as searching for $\mathcal{A}(\widehat{\theta}^{t-1}, \mathbf{x})$ by the definition can be computationally expensive. Our practical strategy is to use the gradient of $\widehat{\theta}^{t-1}$ to guide the selection of the features. Intuitively, we argue that the the features with larger absolute values of $\partial l(\theta^{t-1}, \mathbf{x}, \mathbf{y})/\partial \theta^{t-1}$ are the features $\widehat{\theta}^{t-1}$ relies on.

Finally, we consider the features with values greater than a threshold $\tau(\rho, \mathbf{g})\}$ are the features that are in $\mathcal{A}(\widehat{\theta}^{t-1}, \mathbf{x})$. The threshold hold is set as the ρ^{th} quantile of all the calculated gradients for this sample. The algorithm is shown in Algorithm 1

Algorithm 1: Learning Robust Models with Minimum Supervision

```
Result: $\theta^T$
Input: $T$, $\rho$, $(\mathbf{X}, \mathbf{Y})$;
initialize $\theta^0$, $t = 1$, $\eta$;
while $t \leq T$ do
    for sample $(\mathbf{x}, \mathbf{y})$ do
        calculate the gradient $\mathbf{g} = \partial l(\theta^{t-1}, \mathbf{x}, \mathbf{y}) / \partial \theta^{t-1}$;
        set the threshold $\tau(\rho, \mathbf{g})$ to be the $\rho^{\text{th}}$ quantile of $|\mathbf{g}|$;
        set $\mathcal{A}(\theta^{t-1}, \mathbf{x}) = \{i || \mathbf{g}_i| \geq \tau(\rho, \mathbf{g})\}$;
        sample $\mathbf{x}'$ where $\mathbf{x}'_{\mathcal{A}(\theta^{t-1}, \mathbf{x})} \in \mathcal{X}_{\mathcal{A}(\theta^{t-1}, \mathbf{x})}$;
        calculate the gradient $\mathbf{g}' = \partial l(\theta^{t-1}, \mathbf{x}', \mathbf{y}) / \partial \theta^{t-1}$;
        update the model $\theta^t = \theta^{t-1} - \eta \mathbf{g}'$
    end
end
```

B PROOFS OF THEORETICAL DISCUSSIONS

B.1 LEMMA B.1 AND PROOF

Lemma B.1. *With sample $(\mathbf{x}, \mathbf{y})$ and two labeling functions $f_1(\mathbf{x}) = f_2(\mathbf{x}) = \mathbf{y}$, for an estimated $\theta \in \Theta$, if $\theta(\mathbf{x}) = \mathbf{y}$, then with* ***A3*** *and* ***A4****, we have*

$$d_{\mathbf{x}}(\theta, f_1) = 1 \iff r(\theta, \mathcal{A}(f_2, \mathbf{x})) = 1 \tag{18}$$

$\mathbf{x}_{\mathcal{A}(f,\mathbf{x})} \in \mathcal{X}_{\mathcal{A}(f,\mathbf{x})}$ *denotes that the features of* $\mathbf{x}$ *indexed by* $\mathcal{A}(f, \mathbf{x})$ *are searched in the entire space.*

Proof. If $\theta(\mathbf{x}) = \mathbf{y}$ and $d_{\mathbf{x}}(\theta, f_1) = 1$, according to **A4**, we have $d_{\mathbf{x}}(\theta, f_2) = 0$.

First, we consider one direction $d_{\mathbf{x}}(\theta, f_1) = 1 \implies r(\theta, \mathcal{A}(f_2, \mathbf{x})) = 1$ and we prove this by contradiction.

If the conclusion does not hold, $r(\theta, \mathcal{A}(f_2, \mathbf{x})) = 0$, which means

$$\max_{\mathbf{x}_{\mathcal{A}(f_2,\mathbf{x})} \in \mathcal{X}_{\mathcal{A}(f_2,\mathbf{x})}} |\theta(\mathbf{x}) - \mathbf{y}| = 0 \tag{19}$$

Together with $d_{\mathbf{x}}(\theta, f_2) = 0$, which means

$$\max_{\mathbf{z} \in \mathcal{X}: \mathbf{z}_{\mathcal{A}(f_2,\mathbf{x})} = \mathbf{x}_{\mathcal{A}(f_2,\mathbf{x})}} |\theta(\mathbf{z}) - \mathbf{y}| = 0, \tag{20}$$

we will have

$$\max_{\mathbf{x} \in \mathcal{X}} |\theta(\mathbf{x}) - \mathbf{y}| = 0, \tag{21}$$

which is $\theta(\mathbf{x}) = \mathbf{y}$ for any $\mathbf{x} \in \mathbf{P}$.

This contradicts with the premises in **A4** (θ is not a constant function).

Second, we consider the other direction $r(\theta, \mathcal{A}(f_2, \mathbf{x})) = 1 \implies d_{\mathbf{x}}(\theta, f_1) = 1$ and we prove this by showing its contrapositive proposition holds. (Its contrapositive proposition is $d_{\mathbf{x}}(\theta, f_1) = 0 \implies r(\theta, \mathcal{A}(f_2, \mathbf{x})) = 0$, because, by definitions, r and d can only be evaluated as 0 or 1).

Because of **A3** ($\mathcal{A}(f_1, \mathbf{x}) \cap \mathcal{A}(f_2, \mathbf{x}) = \emptyset$), we have $d_{\mathbf{x}}(\theta, f_1) \geq r(\theta, \mathcal{A}(f_2, \mathbf{x}))$, thus the contrapositive proposition can be shown trivially. □

B.2 THEOREM 3.1 AND PROOF

Theorem. *With Assumptions* ***A1-A4****, with probability as least $1 - \delta$, we have*

$$\epsilon_{\mathbf{P}_t}(\theta) \leq \widehat{\epsilon}_{\mathbf{P}_s}(\theta) + c(\theta) + \phi(|\Theta|, n, \delta) \tag{22}$$

where $c(\theta) = \dfrac{1}{n} \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = \mathbf{y}] r(\theta, \mathcal{A}(f_p, \mathbf{x}))$.

Proof.

$$\widehat{\epsilon}_{\mathbf{P}_s}(\theta) = \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} |\theta(\mathbf{x}) - f(\mathbf{x})| \tag{23}$$

$$= 1 - \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \big(\mathbb{I}[\theta(\mathbf{x}) = f(\mathbf{x})]\big) \tag{24}$$

$$= 1 - \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \big(\mathbb{I}[\theta(\mathbf{x}) = f(\mathbf{x})]\mathbb{I}[d_{\mathbf{x}}(\theta, f_d) = 0] + \mathbb{I}[\theta(\mathbf{x}) = f(\mathbf{x})]\mathbb{I}[d_{\mathbf{x}}(\theta, f_d) = 1]\big) \tag{25}$$

$$= 1 - \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \big(\mathbb{I}[\theta(\mathbf{x}) = f(\mathbf{x})]\mathbb{I}[d_{\mathbf{x}}(\theta, f_d) = 0]\big) - \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = f(\mathbf{x})]\mathbb{I}[d_{\mathbf{x}}(\theta, f_d) = 1] \tag{26}$$

$$= \widehat{\epsilon}_d(\theta) - \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = f(\mathbf{x})] r(\theta, \mathcal{A}(f_p, \mathbf{x})), \tag{27}$$

where the last line used Lemma B.1.

Thus, we have

$$\widehat{\epsilon}_d(\theta) = \widehat{\epsilon}(\theta) + \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{1}[\theta(\mathbf{x}) = f(\mathbf{x})] r(\theta, \mathcal{A}(f_p, \mathbf{x})) \tag{28}$$

where

$$\widehat{\epsilon}_d(\theta) = 1 - \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \big(\mathbb{1}[\theta(\mathbf{x}) = f(\mathbf{x})] \mathbb{1}[d_{\mathbf{x}}(\theta, f_d) = 0]\big), \tag{29}$$

which describes the correctly predicted terms that θ functions the same as f_d and all the wrongly predicted terms. Therefore, conventional generalization analysis through uniform convergence applies, and we have

$$\epsilon_{\mathbf{P}_t}(\theta) \leq \widehat{\epsilon}_d(\theta) + \phi(|\Theta|, n, \delta) \tag{30}$$

Thus, we have:

$$\epsilon_{\mathbf{P}_t}(\theta) \leq \widehat{\epsilon}_{\mathbf{P}_s}(\theta) + \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{1}[\theta(\mathbf{x}) = \mathbf{y}] r(\theta, \mathcal{A}(f_p, \mathbf{x})) + \phi(|\Theta|, n, \delta) \tag{31}$$

□

B.3 THEOREM 3.2 AND PROOF

Theorem. *With Assumptions **A2-A5**, and if* $1 - f_d \in \Theta$*, we have*

$$c(\theta) \leq D_{\Theta}(\mathbf{P}_s, \mathbf{P}_t) + \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{1}[\theta(\mathbf{x}) = \mathbf{y}] r(\theta, \mathcal{A}(f_p, \mathbf{x})) \tag{32}$$

where $c(\theta) = \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{1}[\theta(\mathbf{x}) = \mathbf{y}] r(\theta, \mathcal{A}(f_p, \mathbf{x}))$ *and* $D_{\Theta}(\mathbf{P}_s, \mathbf{P}_t)$ *is defined as in (8).*

Proof. By definition, $g(\mathbf{x}) \in \Theta \Delta \Theta \iff g(\mathbf{x}) = \theta(\mathbf{x}) \oplus \theta'(\mathbf{x})$ for some $\theta, \theta' \in \Theta$, together with Lemma 2 and Lemma 3 of (Ben-David et al., 2010), we have

$$\begin{aligned}
D_{\Theta}(\mathbf{P}_s, \mathbf{P}_t) =& \frac{1}{n} \max_{\theta, \theta' \in \Theta} \Big| \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} |\theta(\mathbf{x}) - \theta'(\mathbf{x})| - \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} |\theta(\mathbf{x}) - \theta'(\mathbf{x})| \Big| && (33)\\
\geq& \frac{1}{n} \Big| \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} |\theta(\mathbf{x}) - f_z(\mathbf{x})| - \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} |\theta(\mathbf{x}) - f_z(\mathbf{x})| \Big| && (34)\\
=& \frac{1}{n} \Big| \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{1}[\theta(\mathbf{x}) = \mathbf{y}] - \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{1}[\theta(\mathbf{x}) = \mathbf{y}] \Big| && (35)\\
=& \frac{1}{n} \Big| \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{1}[\theta(\mathbf{x}) = \mathbf{y}] \mathbb{1}[r(\theta, \mathcal{A}(f_p, \mathbf{x})) = 1] - \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{1}[\theta(\mathbf{x}) = \mathbf{y}] \mathbb{1}[r(\theta, \mathcal{A}(f_p, \mathbf{x})) = 1] && (36)\\
&+ \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{1}[\theta(\mathbf{x}) = \mathbf{y}] \mathbb{1}[r(\theta, \mathcal{A}(f_p, \mathbf{x})) = 0] - \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{1}[\theta(\mathbf{x}) = \mathbf{y}] \mathbb{1}[r(\theta, \mathcal{A}(f_p, \mathbf{x})) = 0] \Big| && (37)\\
=& \frac{1}{n} \Big| \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{1}[\theta(\mathbf{x}) = \mathbf{y}] r(\theta, \mathcal{A}(f_p, \mathbf{x})) - \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{1}[\theta(\mathbf{x}) = \mathbf{y}] r(\theta, \mathcal{A}(f_p, \mathbf{x})) \Big| && (38)\\
\geq& c(\theta) - \sum_{(\mathbf{x},\mathbf{y}) \in (\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{1}[\theta(\mathbf{x}) = \mathbf{y}] r(\theta, \mathcal{A}(f_p, \mathbf{x})) && (39)
\end{aligned}$$

First line: see Lemma 2 and Lemma 3 of (Ben-David et al., 2010).

Second line: if $1 - f_d \in \Theta$, and we use f_z to denote $1 - f_d$.

Fifth line is a result of using that fact that

$$\sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = \mathbf{y}]\mathbb{I}[r(\theta, \mathcal{A}(f_p, \mathbf{x})) = 0] = \sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{I}[\theta(\mathbf{x}) = \mathbf{y}]\mathbb{I}[r(\theta, \mathcal{A}(f_p, \mathbf{x})) = 0] \tag{40}$$

as a result of our assumptions. Now we present the details of this argument:

According to **A4**, if $\theta(\mathbf{x}) = \mathbf{y}$, $d_x(\theta, f_d)d_x(\theta, f_p) = 0$. Since $r(\theta, \mathcal{A}(f_p, \mathbf{x})) = 0$, $d_x(\theta, f_p)$ cannot be 0 unless θ is a constant mapping that maps every sample to 0 (which will contradicts **A4**). Thus, we have $d_x(\theta, f_d) = 0$.

Therefore, we can rewrite the left-hand term following

$$\sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = \mathbf{y}]\mathbb{I}[r(\theta, \mathcal{A}(f_p, \mathbf{x})) = 0] = \sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = \mathbf{y}]\mathbb{I}[d_x(\theta, f_d) = 0] \tag{41}$$

and similarly

$$\sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{I}[\theta(\mathbf{x}) = \mathbf{y}]\mathbb{I}[r(\theta, \mathcal{A}(f_p, \mathbf{x})) = 0] = \sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{I}[\theta(\mathbf{x}) = \mathbf{y}]\mathbb{I}[d_x(\theta, f_d) = 0] \tag{42}$$

We recap the definition of $d_x(\cdot, \cdot)$, thus $d_x(\theta, f_d) = 0$ means

$$d_{\mathbf{x}}(\theta, f_d) = \max_{\mathbf{z}\in\mathcal{X}:\mathbf{z}_{\mathcal{A}(f,\mathbf{x})}=\mathbf{x}_{\mathcal{A}(f_d,\mathbf{x})}} |\theta(\mathbf{z}) - f_d(\mathbf{z})| = 0 \tag{43}$$

Therefore $d_x(\theta, f_d) = 0$ implies $\mathbb{I}(\theta(\mathbf{x}) = \mathbf{y})$, and

$$|\theta(\mathbf{z}) - f_d(\mathbf{z})| = 0 \quad \forall \quad \mathbf{z}_{\mathcal{A}(f_d,\mathbf{x})} = \mathbf{x}_{\mathcal{A}(f_d,\mathbf{x})} \tag{44}$$

Therefore, we can continue to rewrite the left-hand term following

$$\sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) = \mathbf{y}]\mathbb{I}[d_x(\theta, f_d) = 0] = \sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{z}) - f_d(\mathbf{z})] = \sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) - f_d(\mathbf{x})] \tag{45}$$

and similarly

$$\sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{I}[\theta(\mathbf{x}) = \mathbf{y}]\mathbb{I}[d_x(\theta, f_d) = 0] = \sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{I}[\theta(\mathbf{z}) - f_d(\mathbf{z})] \tag{46}$$

where $\mathbf{z}$ denotes any $\mathbf{z} \in \mathcal{X}$ and $\mathbf{z}_{\mathcal{A}(f_d,\mathbf{x})} = \mathbf{x}_{\mathcal{A}(f_d,\mathbf{x})}$.

Further, because of **A5**, we have

$$\sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_t}} \mathbb{I}[\theta(\mathbf{z}) - f_d(\mathbf{z})] = \sum_{(\mathbf{x},\mathbf{y})\in(\mathbf{X},\mathbf{Y})_{\mathbf{P}_s}} \mathbb{I}[\theta(\mathbf{x}) - f_d(\mathbf{x})]. \tag{47}$$

Thus, we showed the (40) holds and conclude our proof.

□

C ADDITIONAL EXPERIMENTS

C.1 THEORETICAL SUPPORTING EXPERIMENTS

Synthetic Data with Spurious Correlation We extend the setup in Figure 1 to generate the synthetic dataset to test our methods. We study a binary classification problem over the data with n samples and p features, denoted as $\mathbf{X} \in \mathcal{R}^{n \times p}$. For every training and validation sample i, we generate feature j as following:

$$\mathbf{X}_j^{(i)} \sim \begin{cases} N(0,1) & \text{if } 1 \leq j \leq 3p/4 \\ N(1,1) & \text{if } 3p/4 < j \leq p, \text{ and } y^{(i)} = 1, \quad \text{w.p. } \rho \\ N(-1,1) & \text{if } 3p/4 < j \leq p, \text{ and } y^{(i)} = 0, \quad \text{w.p. } \rho \\ N(0,1) & \text{if } 3p/4 < j \leq p, \quad \text{w.p. } 1-\rho \end{cases},$$

In contrast, testing data are simply sampled with $\mathbf{x}_j^{(i)} \sim N(0,1)$.

To generate the label for training, validation, and test data, we sample two effect size vectors $\beta_1 \in \mathcal{R}^{p/4}$ and $\beta_2 \in \mathcal{R}^{p/4}$ whose each coefficient is sampled from a Normal distribution. We then generate two intermediate variables:

$$\mathbf{c}_1^{(i)} = \mathbf{X}_{1,2,\dots,p/4}^{(i)}\beta_1 \quad \text{and} \quad \mathbf{c}_2^{(i)} = \mathbf{X}_{1,2,\dots,p/4}^{(i)}\beta_2$$

Then we transform these continuous intermediate variables into binary intermediate variables via Bernoulli sampling with the outcome of the inverse logit function ($g^{-1}(\cdot)$) over current responses, *i.e.*,

$$\mathbf{r}_1^{(i)} = \text{Ber}(g^{-1}(\mathbf{c}_1^{(i)})) \quad \text{and} \quad \mathbf{r}_2^{(i)} = \text{Ber}(g^{-1}(\mathbf{c}_2^{(i)}))$$

Finally, the label for sample i is determined as $\mathbf{y}^{(i)} = \mathbb{1}(\mathbf{r}_1^{(i)} = \mathbf{r}_2^{(i)})$, where $\mathbb{1}$ is the function that returns 1 if the condition holds and 0 otherwise.

Intuitively, we create a dataset of p features, half of the features are generalizable across train, validation and test datasets through a non-linear decision boundary, one-forth of the features are independent of the label, and the remaining features are spuriously correlated features: these features are correlated with the labels in train and validation set, but independent with the label in test dataset. There are about $c\dot{n}$ train and validation samples have the correlated features.

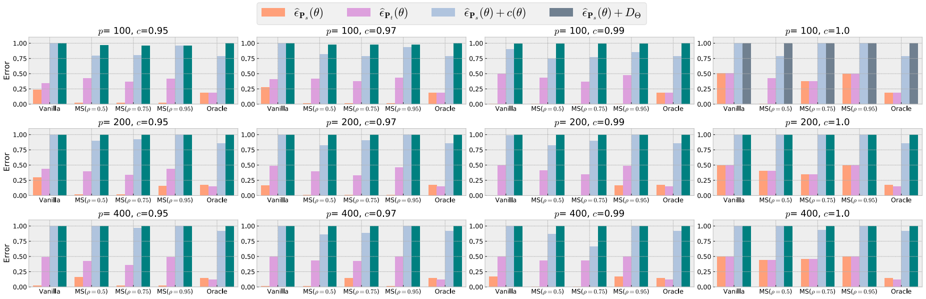

Figure 3: Results of Synthetic Data with Spurious Correlation. Each panel represents one setting. Five methods are reported in each panel. For each method, four bars are plotted: from left to right, $\widehat{\epsilon}_{\mathbf{P}_s}(\theta)$, $\widehat{\epsilon}_{\mathbf{P}_t}(\theta)$, $\widehat{\epsilon}_{\mathbf{P}_s}(\theta) + c(\theta)$, and $\widehat{\epsilon}_{\mathbf{P}_s}(\theta) + D_\Theta$.

Results are reported in Figure 3, where each setup we ran 3 random seeds and report the mean and standard deviation. We train a vanilla method, minimum supervision method with different hyperparamter ρ, and an oracle method that uses data augmentation to randomized the previously known spurious features. The results show the advantage of the new method consistently, although still not compared to the method with prior knowledge. We also calculate the $c(\theta)$ as we perform adversarial attacks over the spuriously correlated feature space, we also calculate D_Θ as defined in (8). We compared $\widehat{\epsilon}_{\mathbf{P}_s}(\theta) + c(\theta)$ and $\widehat{\epsilon}_{\mathbf{P}_s}(\theta) + D_\Theta$ and the results suggest that clearly $c(\theta)$ offers a more accurate assessment of the target error than D_Θ.

C.2 REAL IMAGE CLASSIFICATION: MORE DETAILS

The main experiment setup follows the setup of (Bahng et al., 2019), and the setup can be conveniently replicated by the GitHub repo associated with the paper (Bahng et al., 2019). Although results of ImageNet-C are also reporeted by (Bahng et al., 2019), their github repo does not provide the corresponding replication scripts, so we also skip the information. Additionally, we report another ImageNet level test set that is independently collected, and has only sketch images.

We rename the "bias" and "unbiased" in (Bahng et al., 2019) to "standard accuracy" and "weighted accuracy" to align the terms we use in this paper and also help to explain the results. Intuitively, "weighted accuracy" refers to the evaluation mechanism that the test samples with unusual texture will have more weights.

Again, following the setup in (Bahng et al., 2019), the base network is ResNet, and we compare with the vanilla network, and several methods that are designed to reduce the texture bias: including StylisedIN (Geirhos et al., 2019), LearnedMixin (Clark et al., 2019), RUBi (Cadene et al., 2019) and ReBias (Bahng et al., 2019).

Finally, to get the reported performance, our MS method uses an extra heuristic, such as we only optimize (16) for half of the batch, and optimize the other batch with the vanilla training (1). Despite this heuristic used, the main message remains: MS method, as a method that does not use the knowledge of the spurious correlated features, can compete with the methods that use the knowledge explicitly.